SUSMAN GODFREY L.L.P.

November 18, 2024

Hon. Ona T. Wang
United States Magistrate Judge
Southern District of New York

Re: *The New York Times Company v. Microsoft Corporation, et al.*,
Case No.: 23-cv-11195-SHS-OTW

Dear Magistrate Judge Wang:

Plaintiff The New York Times Company ("The Times") seeks an order compelling OpenAI to produce documents in response to the following Requests for Production:[1]

RFP 2: This Request seeks documents Defendants "have gathered to submit or have submitted to any legislative or executive agency, committee, or other governmental entity in the United States that concern or relate to the allegations in the Complaint."

For the purposes of this motion, The Times specifically seeks an order that (i) OpenAI identify any relevant proceedings in which it has produced documents related to the issues in this case, which will facilitate additional negotiations about the appropriate scope of OpenAI's response to this RFP; and (ii) OpenAI produce Interrogatory responses 2, 3, 5, 6, 7, and 10 that it submitted to the FTC for Civil Investigative Demand No. 232-3044. *See* Ex. 2 (copy of FTC Interrogatories). These responses will yield information relevant to this case regarding the relationship between the various OpenAI entities (FTC Rog 2), OpenAI's employees (FTC Rog 3), OpenAI's revenues (FTC Rog 5 and Rog 6), OpenAI's licensing agreements (FTC Rog 7), and information about third-party use of OpenAI's large language models (FTC Rog 10).[2] *See* Ex. 2 at 3-4. There is no burden to OpenAI producing these materials, and it should do so. *See, e.g., Waldman v. Wachovia Corp.*, 2009 WL 86763, at *1-2 (S.D.N.Y. Jan. 12, 2009) (ordering production of materials provided to regulators because the "burden is slight when a defendant has already found, reviewed and organized the documents"). OpenAI argues that some of the information in these responses may be duplicative to information being sought in other discovery requests, but The Times cannot evaluate that assertion without first reviewing the Interrogatory responses. The parties can later evaluate whether these documents moot or narrow any outstanding discovery issues.

RFP 17: This Request seeks documents "concerning the compilation, acquisition, and curation of Training Datasets," including "communications concerning the relevance and impact

[1] The parties met and conferred by videoconference on November 14 and November 18. A November 8, 2024 letter detailing The Times's positions on these RFPs is attached as Exhibit 1. Copies of email correspondence between the parties is attached as Exhibit 5. The Times's First Set of RFPs (Requests 1-15), and OpenAI's Responses, have been filed on the Docket at Dkts. 128-1 and 128-2. The Times's Second Set of RFPs (Requests 16-72), and OpenAI's Responses, have been filed on the Docket at Dkts. 283-1 and 283-3. The Times's Third Set of RFPs (Requests 73-95), and OpenAI's Responses, have been filed on the Docket at Dkts. 283-2 and 283-4.

[2] To resolve a dispute in the *Authors Guild* case, OpenAI produced responses to some of these Interrogatories, but not the ones requested in this motion. *See Authors Guild*, Case 23-cv-8292, Dkt. 106.

of paywalls on building Training Datasets or gathering training data." There is one remaining and narrow dispute: whether OpenAI will produce documents concerning the "impact" of paywalls on training data, which is broader than OpenAI's proposal to limit productions to documents concerning OpenAI's "approach to" paywalls. The broader language is important to ensure that OpenAI does not withhold documents addressing how paywalls impact its compilation of training data on the alleged ground that the document does not address OpenAI's specific "approach" to paywalls. If OpenAI confirms that its proposal will not exclude documents concerning the "impact" of paywalls, then there is no dispute, but OpenAI has not yet done so.

RFP 24: This request seeks documents "concerning any analyses, studies, measurements, testing, experimentation, assessments, or other evaluation of the incremental value of including Journalism content, including Times Content, in Defendants' Training Datasets." The use of journalism content to train Defendants' generative AI models is a central issue in this litigation case, and documents reflecting the value of news content are relevant to both fair use and damages.

OpenAI has agreed to produce documents discussing *OpenAI's* evaluation or assessment of the value of including Journalism content in the text training datasets used to train the relevant models but refuses to produce documents concerning evaluations or assessments prepared by *entities other than OpenAI*, including any evaluations or assessments prepared by Microsoft or by academic researchers. This is improper. Any such documents in OpenAI's possession, custody, or control concerning the value of journalism for generative AI training datasets are relevant to this litigation and should be easy to produce through the normal discovery process.

RFP 49: This request seeks documents "concerning Defendants' monitoring, tracking, and knowledge of investigations into whether Defendants' AI Models contain copyrighted content, including Times Content." Defendants' knowledge of investigations into whether their models contain copyrighted publisher content is relevant to willfulness.

OpenAI has agreed to produce documents concerning whether the relevant models were trained using articles published by The Times, Daily News Plaintiffs, and the Center for Investigative Reporting, but objects that it does not understand what The Times means by "copyrighted publisher content." While The Times maintains that this phrasing is clear on its face, The Times has offered to use the defined term "Journalism" content to clarify its request. The Times defines "Journalism" as "the activity of writing or creating content for newspapers, magazines, news websites, mobile applications, television, podcasts, or any other publication and/or news outlet, and includes the work of The Times as alleged in the Complaint." Dkt. 283-2 at 3. OpenAI has not yet responded to The Times's offer, as of this filing. The Times appropriately seeks documents broader than those that expressly mention the three plaintiffs in this case. Otherwise, for example, OpenAI could withhold a document concerning an employee's knowledge of their improper reliance on copyrighted Journalism content on the ground that the document did not specifically address one of the three plaintiffs in this case.

RFP 86–87: RFP 86 seeks documents concerning Defendants' use of generative AI models to create synthetic datasets, and RFP 87 seeks documents concerning Defendants' use or contemplated use of synthetic datasets, including to train their generative AI models. "Synthetic" data refers to "artificial, algorithmically-manufactured data, including data generated using a generative AI model." Dkt. 283-2 at 4. Defendants' use of synthetic datasets created by the outputs of generative AI models that are trained or fine-tuned on Times Content are relevant to The Time's copyright infringement claims—any synthetic datasets created by OpenAI using Times content

may be infringing works, and OpenAI's use of Times content to create synthetic datasets is also relevant to fair use because these datasets may impact the market for training data.

With respect to RFP 86, OpenAI seeks to improperly limit productions to Times content, as opposed to Journalism content more broadly. With respect to RFP 87, OpenAI has refused to produce documents concerning its "contemplated use" of synthetic datasets, which is improper because that would exclude, among other things, documents concerning how OpenAI has tried (unsuccessfully) to use synthetic data (which would then help show the value of authentic journalism content for training). Finally, for both RFPs, OpenAI seeks to limit productions to a "sufficient to show" production, which is insufficient given that these RFPs address core disputed issues in this case.

RFP 95: This request seeks "Documents and communications concerning [redacted] OpenAI should collect and review responsive documents and log any communications over which it seeks to claim privilege. Indeed, OpenAI has never provided any information about the volume of documents at issue, undermining any suggestion that the privilege review process would be unduly burdensome.

Respectfully submitted,

/s/ Ian B. Crosby
Ian B. Crosby

cc: All Counsel of Record (via ECF)